**Google Store Application Executive Summary**

**Introduction**

The data has been collected on June 2019 it has thirteen features, 10841 observation and the missing values about 1487 from 3 features Rating, feature has 14% .

**Objective:**

To find out which are the most independent features that can at best describe the dependent feature (Rating).

**Methodology:**

google store data set csv file was loaded into panda library dataframe structure for manipulation, cleaning, and modeling data has values were numbers, character and some special character like $, +, - ‘’

regex and lambda for remove all special characters to make the features types either categorical or numeric

* android version, android current version and type were their missing values filled up using the mode which the most frequent values in each feature
* seanborn were used for plotting the and check the correlation between variables and the target feature
* plotting and analysis were accomplished on data with Plotty Express for statically

graphs numpy library were used for.

* Engineered new features sales, year, month for more and backward elimination were applied with regression where the p greater than five would be dropped.

**Modeling**

The target variable is categorized on four groups low good, medium, and very good

Select best the best model result for prediction

|  |  |
| --- | --- |
| **MODEL** | **RESULT** |
| KNeighborsClassifier | 0.7206765728738539 |
| LogisticRegression | 0.7031710914454278 |
| KNeighborsClassifier | 0.68547197640118 |

**Conclusions & Recommendations**

The most application category which are Family, Tools, games and Life style theses what should the developer focus on which are the most

The application size has zero impact when to application development doesn’t matter

92% of the share market are free applications however 8% make almost 4,000,000,000

Google launch first up for sale on 2011

Finance and Tools are the most reviewed

The column features Category, Content Rating, type, are describes and will predict the rating with 70 to 72 confidence interval